如何做出道德的人工智能体？心理学的视角

喻丰 许丽颖
西安交通大学人文社会科学学院
通讯作者：喻丰，yufengx@xjtu.edu.cn

摘要：人工智能的飞速发展产生了一系列道德困境，如何做出道德的人工智能体（如道德人工智能机器人）成为了必须回答的问题。本文从心理学视角探究了人工智能体是否可能被赋予道德地位、被如何赋予何种道德地位；人工智能体是否需要及需要何种道德能力；人工智能体如何获得及获得何种道德规则；人工智能体能够如何深化人类对于人性、关系以及多样化的道德理解等多种问题进行了回答。从心理学的理论和实证研究出发，切实回答如何做出道德的人工智能体的疑问，期望对以人为中心的人工智能研究提供道德心理学智慧。

关键词：人工智能；道德；道德判断；人机交互；人—机器人交互

一、人工智能体的道德地位（moral standing）

（一）人工智能体需要道德地位吗？

日常生活中，无论是朴素民众（folk people）还是处于人工智能研究顶尖浪潮上的科学家，都询问出类似问题："如何做一个道德的机器或机器人或人工智能体？"（虽然人工智能可能只是底层的计算或者是代码，但是本文所指人工智能体包括人工智能程序、人工智能应用、人工智能机器、人工智能机器人等，我们在本文中采用"人工智能体"来代表所有这些人工智能实体）。但从语言学上分析，当他们问出这样的问题时，他们已然承认机器、机器人、或者人工智能体具有道德地位。亦即，人类普遍有一种想要赋予人工智能体道德地位的朴素倾向，尤其是，当人类知觉、注意并意识到人工智能体其智能（intelligence）或者其进行具体事务的能力（competence）要远胜于自己时，人们更可能赋予人工智能体道德地位（Gray, Gray, & Wegner, 2007）。作为美好的愿景，我们希望那些智能和能力远胜于人类的人工智能体，它可以是道德的。

这样说，便表明人类在心理上是存疑的，即人类害怕、恐惧、忧虑、担心人工智能体是不道德的。我们无法排除这种可能，出于对人类物种的中心主义和自尊，人类通常情况下会在除去智能和能力的其他方面对人工智能体进行贬损。从社会认知观的角度上来看，人类认识世界通过两个基本维度：一曰能力（即我们这里所讨论的智能和处理具体事务的能力），二曰温暖（即进行良好社会交往和道德的水平）（Cuddy, Fiske, & Glick, 2008）。如果能力上不能贬损人工智能体，那么人类倾向于在温暖上贬损它们，最常见的是倾向于认为它们没有道德地位。

这是这个阶段人类认识上较为可笑的矛盾。他们一方面外显地认为机器、机器人或者人工智能体没有道德地位，另一方面他们又内隐地、悄无声息地、不由自主地、未经控制地存有将人工智能体作为一个具有道德地位的道德实体的判断。试想一下如下场景：一位正在打字，写着学术论文的哲学教授，一边写下类似："人工智能无论如何强大、如何智慧、如何有能力，都改变不了其所属的工具地位，它永远不可能成为承担道德责任的主体，也永远不可能享有道德权利与义务，总而言之，它不具有道德地位"。而这时他的孩子被他家的扫地机器人飞驰而过撞伤了腿，他扔下论文，抄起扫地机器人便砸在地上，忿忿不平地骂道："讨厌的死机器人、坏机器人，我一定要惩罚你"（王银春, 2018）。这样的情境不应该只是笔者臆想出来的，笔者相信读者但凡读到此处，脑中都有类似场景鲜活的画面感，而不会觉得陌生。举这个例子就是想说明，在人类认识上对于人工智能体其道德地位存在内隐和外显的分

离。

人类对人工智能体在道德地位上存在认识的内隐外显分离当然不是一件好事，解决分离的途径便是达成一致。内隐和外显并不是一个连续维度上的两个极，它就是一个二分的、要么内隐要么外显的维度，因此若要达成内隐和外显认识上的统一，人们要么在认识上取消人工智能体的道德地位，要么在认识上赋予人工智能体道德地位。究竟是取消还是赋予它们道德地位，这取决于人工智能体是否可能作为一个道德的实体（Entity）而主动地、自发地、有意图地、能动性地去进行具有道德意涵的行为。在现在这个时代和阶段里，它们似乎不可能，即使它们做出了某种行为，而这种行为又伤害了其他生命体，它们在人们的推断中也不可能是主动的、有意图的、自发的、自为的、受自己控制的。因为现阶段的人工智能体并未发展到机器或机器人能够有意图地自发行动的程度。就这个时代这样的人工智能发展水平来取消人工智能的道德地位是合理的。但人工智能的发展是否永远不可能发展出有意图的、自主的道德机器人，这是存在疑问的。哪怕有极小的可能，某一天一个能够自由决定其行为的人工智能体出现，取消它的道德地位都是不可以的。从这个意义上说，对人工智能体的道德地位至少持有一种相当保守的态度，即哪怕现在其不具有道德地位的先在假设，而赋予其道德地位以匹配人类内隐的认知观，这都是合理的。

（二）如何赋予人工智能体道德地位？

假使我们赋予了人工智能体以道德地位，那我们便必须操作化地来说，在具体情境中它会与没有被赋予道德地位的人工智能体有何不同。当然，从推论上来说，这需要人类在心灵上假设人工智能体至少是有意图的。在具体行为层面上，人们则会给被赋予了道德地位的人工智能体以道德权利和义务。同时，其还必须承担道德责任。换句话说，即人类将这种人工智能体纳入人类的道德圈（moral circle; Singer, 2001; 喻丰，许丽颖, 2018）中，给予人工智能体以道德考量，关心人工智能体作为一个行为发出者或者一个行为的接受者的道德意涵。

一方面，人工智能体需要获得道德权利和义务。几年前，谷歌机器狗被研究人员或是普通民众用脚踢的视频在网络上被大量传播，中国也有研究人员号称做出了机器狗，在测试其行走稳定程度时，其方式也是用脚踹它的身体，而它在踉跄之后还能重新保持平衡，继续行走。这样的视频得以传播的原因并非是人们赞叹人工智能技术的先进发展，而是人们对机器狗产生了恻隐之心，并对研究人员产生谴责或谴责的倾向。这体现了前面所述人类对人工智能体内隐的道德立场，不自觉地赋予了人工智能体以道德权利，把人工智能体当成了有机生命体来看待。机器狗被当成有机生命体是可能的，因为其毕竟有大量拟人化的成分在，其四肢和躯干、其行动模式都在拟人（Phillips, Zhao, Ullman, & Malle, 2018）。我们这里说的拟人不仅仅指人，还包括动物等其他生命体。试想上文所说的扫地机器人，它会内隐地被赋予道德权利吗？似乎不会，但我们却可以把它当成道德对象来谴责。也就是说，扫地机器人至少在现在这个发展阶段我们不会关心其情绪、情感、心理状态，我们不会因为其所处的痛苦而悲愤异常、伤心流泪。主要原因是扫地机器人只是一个圆盘形状、其行动轨迹也与其他生命体大相径庭，甚至看起来普通民众并不能理解它是一个人工智能机器，而仅仅把它当机器。若要赋予人工智能体以道德地位，则其道德权利和义务无论在任何情况下，哪怕其外观不似人类，哪怕其内在只是算法，我们都应记住它们可以是道德的对象。我们理应以道德圈内之物对其施加道德考量，对其受到的伤害和痛苦予以减少、痛感情绪。

另一方面，人工智能体若被赋予道德地位，其也应该担负道德责任。研究发现，若自动驾驶汽车处于一个道德考量中，面临一种需要杀一救五或者杀一救多，而杀的这个“一”是自动驾驶汽车的主人的话，人们只会将很少的道德责任归结于这辆无人驾驶汽车，而大多数人认为应该杀一救五，但大多数人都倾向于不去购买这辆无人驾驶汽车（Bonnefon, Shariff, & Rahwan, 2016）。笔者自己的研究也发现，如果一辆智能汽车突然失灵、开始鸣笛，鸣笛的

声音伤害了正从超市出来经过这辆汽车的一个被母亲推在摇篮中的孩子的听力，那么人们在归责时倾向于将更多的道德责任归结为设计这辆智能汽车的程序员、生产商、售卖商甚至这位母亲，而将很少的道德责任归结为是这辆汽车所为（邬家骅, 喻丰, 许丽颖, 2016）。两个实验都发现了同样的效应，即人们实际上很少会将道德责任归结为人工智能体，但值得指出的是，人们并不是完全不把道德责任归结为它们。我们不倾向于给人工智能体归责，是因为在人类归因的规律上，存在一种所谓的基本归因错误（Fundamental Attribution Error, FAE; Ross, 1977），即我们倾向于将行为的原因归结于行为人本身，而非除人之外的其他一切，比如情境。一个被赋予了道德立场的人工智能体必然是能够承受道德责任的行为主体，在认识上人们倾向于让其负少量责任，这也许也因为这个时代的人工智能发展让人不那么容易将其作为道德主体看待，而若要赋予其道德立场，我们需将其作为一种特殊的道德实体来类比于人或动物形成对其进行道德责任归因或者是道德谴责、道德表扬的独特模式。

（三）赋予人工智能体何种道德地位？

上述讨论我们举了两例，分别是一个受虐待的人工智能体和一个需要承担道德责任的人工智能体，用以说明一个具有道德地位的人工智能体至少在人的认识论上如何体现其道德地位。从心智知觉（mind perception）理论来看，道德立场就区分为两种：一者为道德客体（moral patient），一者为道德主体（moral agent）（Gray, Young, & Waytz, 2012）。道德客体是这个道德一对概念中行为的接受者，它引起人类的同情等情感反应，正如上文所说被踢的机器狗一样。道德主体是这个道德一对概念中行为的发出者，它承担行为产生后果所带来的道德责任，正如上文所说损伤了孩子听力的智能汽车一样。

对道德客体所承受痛苦的体验（experience）更好理解。道德涉及伤害（Haidt, 2007）。正如宠物被伤害与孩子被伤害，人所激起的生理机体反应相似，我们推测，若人工智能体受到伤害，人的反应应该在质上与人或动物受到伤害无异，也许存在量上的区别。但这一观点也不绝对，因为人工智能体毕竟不是有机体，即使作为有机体的植物受到伤害（如人砍伐树木、蹂躏花朵），人们对其的反应是否与对人或者对动物相同，这是尚待探讨的问题。更遑论非有机体，研究发现，对于在纯净的雪山上钉钉子这种类似伤害雪山的行为，人们不倾向于将其看做伤害，而倾向于将其看做对纯洁的污染（Frimer, Tell, & Haidt, 2015）。因此作为非有机体的人工智能体如若受到伤害，人们将作何反应首先在于人们将其是否看做人或者类人的生物，当然这取决于人工智能体本身的特征和人们所处的情境。作为道德客体，人们对人工智能体受到伤害的反应似乎还需要更多的研究。

对道德主体所需承担责任的判断要更为复杂。以现在人工智能的发展来看，人工智能体实际上是披着外壳的算法。如果是这样的话，那么这件事情将变得相当复杂。如若这个人工智能体无法自动进行后续的学习和迭代，那么这种人工智能体将保持其出厂时被程序员所设计的样子。这种人工智能体是否会被当成道德主体，人们在考量时理应是存疑的。因为其所进行的行为均由已有程序或者基于其训练库所产生的算法来决定。理论上来说，将责任归结于设计其算法的程序员或者所使用的训练材料似乎更为合理。在这个意义上，这样的人工智能体也许在实践中只能被少量地进行道德责任归因（喻丰, 许丽颖, 2019）。如若这个人工智能体是有自主学习能力的，它可以根据使用者具体使用的经验不断学习迭代（如各种可训练的聊天机器人），那么对其的道德责任归因应该更复杂。比如若聊天机器人在设计时只是可以进行基本的日常会话，但是在经过与各种人大面积的聊天之后，它学会了对女性的歧视。那么这种不道德的行为应该归咎于谁？是设计它的程序员？是那些训练它的、和它聊天的人？还是这个聊天机器人本身？这就如同孩子模仿暴力电影中的情节而杀人，我们会把多大的责任归结于电影呢？当然，这种道德责任归因还需要考虑人工智能体本身的特征（如拟人化程度等）、甚至更为宏大的社会文化背景（如中国人更倾向于向外归因等）（彭凯平, 廖江群,

2009）。但对类似问题的探索至少在现在似乎还没有答案。

二、人工智能体的道德能力（moral competence）

（一）人工智能体是否需要道德能力？

如果人工智能体被赋予了道德立场，那么其是否真的具有可以进行自主道德判断和行为的能力便更为重要。现有的人工智能体应该说都缺乏这种道德判断和进行道德行为的能力。正如图灵测试（Turing test）作为一种判断人工智能的古老标准一样，所谓道德图灵测试（Moral Turing Test, MTT）也被提出用以检验人工智能体是否具有道德能力（Wallach & Allen, 2008）。围绕道德图灵测试的争论很多，笔者不再赘述，就现有观点来看，道德图灵测试并非是一个理论完备、方法可靠、易于操作的客观标准，正如图灵测试都很难真的测试人工智能，道德图灵测试也保留有极强的心理学行为主义刺激与反应联结的观点（Arnold & Scheutz, 2016）。

行为主义观虽然在人工智能领域应用甚广，但从人类心理发生、发展、变化规律的角度看，它极为忽视人类心理具体的运作规律，以现代心理学的角度，它是被抛弃的传统理论。代之以行为主义的无论是精神分析、人本主义还是认知主义，都强调人心具体运作的内部过程。虽然深度学习也采用神经网络模型，但从心理学的角度来看，这是底层的、非心理化地展示人类心理运作黑箱的过程。真正有心理学意蕴的过程并非生物过程，当然，也并非最底层的算法过程，它可能是基于生物和算法之上的心理变量的运作模式。做出道德判断和道德行为必须要求人或者人工智能体有能做出这类判断和行为的能力，同时要求其有做出这类判断和行为的动机与倾向。能力、动机、倾向都是算法和生物过程之上的心理过程。道德能力至少是动机、倾向之前的变量，没有道德能力，无所谓动机或者倾向，因为缺乏能力便缺乏做出道德判断和能力的可能（喻丰, 韩婷婷, 2018）。因此，人工智能体若要满足人类的需求和期望，它必须具有道德能力。

但对人工智能体是否具有道德能力的判断明显不应该通过某种情境看其是否在进行道德判断时能够以假乱真地骗过人类（即所谓的道德图灵测试）。这种测试以先进的心理学观看来更像是抖机灵式的、相对幼稚的寻求捷径。道德能力是否具有，应该看道德能力本身，而不应转而以某种类似预测效度的方式去考量。如果非要以拐弯的方式去考察，那只能是因为无法考察道德能力本身。诚然，道德能力的获得以及道德能力的高低也许有不同的获得方式：自上而下抑或是自下而上。但无论何种方式获得，道德能力应该也是一种或多种实体，可以直接考量（Malle, 2014）。

但需要说明的是，道德能力（moral competence）是否是人工智能体能力（competence）的一种？如果用社会认知的观点来看，道德与能力是相对的，甚至在人的认识论上，某种程度二者还是此消彼长、有种零和感觉的。从这个意义上讲，道德能力看似是个有矛盾的词语。但是将能力之前冠以道德，是在强调能力。以笔者看来，道德能力确是能力的一种，是那些能够用以帮助做出道德判断和道德行为的特定能力，正如功能型的机器人有其功能型的能力（如举重机器人有能承受巨大的重量并将其托起的能力）一样。

但是，一个道德能力强的人工智能体是否一般性的能力都强呢？以及一个一般性能力强的人工智能体是否道德能力强呢？这问题类似心理测量学的信度与效度关系，一者包括于另一者，但并不一定对应增长。从这个时代人工智能的发展浪潮来看，人们倾向于去制造那些一般性能力强或者是功能性能力强的人工智能产品，而忽视其道德能力。人工智能体道德能力的建设要远远困难于其功能性能力。可以试想如若一个一般性能力极强，具有极高智能的机器人缺乏道德能力的设定，其道德能力是在其进行事务和社会活动时所自行习得的，那么它有可能获得在人类看来完全错误的道德规则，产生与人相悖的道德情感，进行非人化的道德判断和行为，这是我们不愿看到的。由于对道德能力的研究无论是其难度还是完成度都要

困难于对人工智能体一般能力的研究，因此笔者建议在人工智能体缺乏道德能力的当下，不应急于追求其一般能力，或曰其智能。

（二）人工智能体需要何种道德能力？

既然人工智能体需要道德能力，那么其道德能力是什么，这涉及到对道德能力的分类。一种广泛的观点认为，道德能力包括以下五种材料：规范系统、道德词汇系统、道德认知和情感、道德决策和行为、道德交流（Malle, 2014）。

第一，道德规范。谁都能理解道德规范的内容，但似乎谁也不能完整地说出道德规范具体是什么、有哪些。在日常生活中，我们根据道德规范来判断一个行为是对或错，但我们也并不清楚这种判断所依据的原因究竟有多少个，它们是如何归类的。如果我们很难清楚这一点，程序员在编程时就很难将这些规则表征为可供机器识别和应用的规条，也无法说明这些规条具体在何种条件下起作用。对于道德规范的探讨，我们还将在后一个部分进行详细的阐述。

第二，道德词汇。道德规范系统必须有语言和算法的表征，之前研究认为，道德词汇系统至少包括三个方面：规范及其属性的词汇（如公平、美德、互惠、诚实、责任、禁止、应该等）、规范违反的词汇（如错误、有罪、鲁莽、窃贼，也包括有意、故意等）、对违反做出反应的词汇（如责备、训斥、原谅、宽恕等）。这些词汇构成了狭义上道德规范表征、广义上道德能力的基本骨架（Malle, 2014）。如研究发现，进行道德批判实际上只需要一个两维度、28 个动词的词语系统。这两个维度是强度和人际，这 28 个词分别是：控诉、指责、批评、责备、反对、挑剔、否决、抨击、严责、攻击、非难、痛斥、指控、诽谤、诋毁、谩骂、数落、呵斥、责骂、严惩、惩戒、警告、申斥、训诫、苛责、侮辱、责难、声讨（Voiklis, Cusimano, & Malle, 2014）。值得说明的是，当我们把这 28 个词翻译成中文时，其意义也许并不一一对应，若用中文词汇来建构道德批判，可能形成的是不尽相同的语词结构。这也提示我们，在道德能力的建构上，仅从语言学角度便存在文化差异。

第三，道德认知和情感。道德认知主要涉及对人工智能体的道德判断，但这种判断似乎可以粗略地首先分为两种：对事件的判断或者对行为主体的判断。对事件的判断包括评价事件、行为、行为结果是好是坏、是错是对、是否可允许等。对行为主体的判断包括评价其是否应负道德责任、是否值得被责备或赞扬（Malle & Scheutz, 2014）。当然，道德认知或曰道德判断离不开情感，这是自休谟以降直至上世纪末、本世纪初才在心理学中复兴的一种情感主义倾向（喻丰, 彭凯平, 韩婷婷, 柴方圆, 柏阳, 2011）。研究发现，人们会直觉地对诸如用国旗擦马桶、吃掉自己家被撞死的狗、兄妹乱伦、答应母亲过分的遗愿却无法完成、和冻鸡发生性行为等事件快速、直觉、无需努力、不假思索地做出其不道德的判断，而无法解释为何（Haidt, Koller, & Dias, 1993; Haidt, 2001）。盖因其做出不道德这种道德判断之前产生了厌恶的情绪。对天桥问题与列车问题的解答不同，也因为二者激活了不同区域的大脑皮层，前者激活了更多与情绪相关的脑区（Greene, Sommerville, Nystrom, Darley, & Cohen, 2001）。

第四，道德决策和行动。道德决策与行动不同于道德判断，人工智能体知善恶、识好歹、明是非，但并不一定真的决定与真的做出良善的行为。通常情况下，人类做出道德决策与行为很大程度上是基于系统一思维的，甚至人类在判断其他人是否道德时也基于系统一思维（Yu & Peng, 2014）。所谓系统一思维，意指那些快速、直觉、不费力气、不加思索、情绪化的思维方式，它用于处理人每日所面临的大部分、大量、扑面而来的繁琐信息；而系统二思维，意指那些缓慢、审慎、耗费努力、仔细加工、理性化的思维方式，它用于处理人在注意指向、空闲而重要时的少量信息（Kahneman & Egan, 2011）。对于人工智能体来说，按现在的运行方式，其行为应该是计算的结果，也就是说，它主要是基于第二思维系统的，且更为理性。由于其能够处理的信息远多于人类，第二思维系统不仅让其能够更好地指导行为，

更有利于设计者更好、更方便地设计足够第二思维系统理性决策的概念表征系统。

第五，道德交流。当人工智能体被赋予道德立场时，它很可能成为被谴责的对象。前面四种道德能力的材料都是个体化的，而人工智能的发展和其道德立场的获得一定会使其身处社会情境中。它会受到谴责，也可能承受非难，它将要辩护，也倾向解释。但无论它在社会交流中对于道德谴责、责备、表扬还是赞颂做出何种辩护、解释、归因，它们都需要这种进行道德交流的能力。道德交流的能力基于前述四种材料之上。

应该说，对这五类道德能力的细致刻画是逐步的、从浅到深的、从个体到社会的、从基础到应用的。

（三）人工智能体的道德能力是否不同？

从上述道德能力的五种材料来看，事实上人类也需要或者也内隐地被设计为具有这五种能力。但人工智能体的道德能力与人类的道德能力相同吗？对这个问题的回答如果是相同，那么我们就在假设人类思维、情感和行为的规律也是基于底层的生物计算。这是一种被许多人持有但未经证实的假说。某种程度上，人类也可能有独立于生物过程与底层计算过程而并不由这些过程决定的高级心理过程。但由于心理学缺乏宏大理论（Grand Theory）的支持，因此我们其实并不清楚也并不明白心理过程究竟是如何铺陈开来的。其道德能力是否不同，只能说很大程度上可能并不相同。

而人与人工智能体道德能力的获得方式可能就完全不同了。从基本原理上推断，人可能生而具有能够学习道德词汇、形成道德规范的语言装置（Chomsky, 1975），但是人的规范系统、道德词汇系统、道德认知和情感、道德决策和行为、道德交流应该都是后天所习得的，人具有习得这些材料，并将其运用的先天能力，但是人缺乏先天出生时就有的这种材料。我们可能有先天转换生成语言装置，也可能有进化而来的适用于学习这些材料的心理基质，但是我们缺少具体学习的内容。具体学习的内容必须是后天习得的，是经验性的。而人工智能体却不同。如果我们能将这些道德能力操作化变成可实现的代码，或者变成可供学习的材料，那么一个具体的人工智能体很可能在其诞生之初便具有了道德能力。但由于人工智能体是不端进行学习和迭代的，那么其所面临的任务与环境中进行的新的学习也能改变或者增加（当然也能减少或者扭转）这种道德能力。用心理学术语来进行类比，出厂设置的道德能力是基于研究和理论的，这似乎是自上而下（top-down）的道德能力；而后期经验性的习得，这似乎是自下而上（bottom-up）的道德能力。前者是人类较少的，因为人类在出生时不具备道德能力的内容，而只有获得道德能力的潜在可能，但人类在后期习得道德规则，进行道德发展后，却能自上而下地去根据道德规则来进行行为；而后者是人类较多地，人类道德发展过程中所经历的从他律到自律或者从前习俗到习俗到后习俗的阶段均缺不了自下而上的经验过程。这种人工智能体获得道德能力方式上与人类的差异便决定了给予人工智能体何种社会经验或者任务是极端重要的，这会导致其道德能力的不同。当然对于其诞生时应该生儿镶嵌何种道德的研究更为基础，也正是现在研究者所进行的。

值得一提的是，经常会有学者反驳说，人工智能永远无法获得真的道德能力，因为人工智能不可能拥有自我、意识、审美、共情、心理理论与观点采择的能力等。而实际上，自我、意识等过程如果也是基于神经生物过程与计算过程的话，那么人工智能体便有拥有这种能力的可能。而审美过程实际上与道德过程类似，最简单都可以采用训练集标注美丑程度的方式轻松让人工智能体习得。共情、心理理论与观点采择则相对麻烦，这标志着人工智能需要去用对方（可能是人、也可能是其他人工智能体）的视角去体会对方的想法与情绪。揣度他人的想法并不困难，人的察言观色、推断意图理论上都能用训练集来进行学习，困难的是，人工智能需要决策是它需要揣度到他人揣度他心理过程的能力。我们都知道曹操败走华容道，而每到一个岔道口，他都会进行选择，在这个选择中他必须揣度诸葛亮是如何判断他的心理，

从而进行抉择。故事中的曹操每做一次选择都会嘲笑孔明少谋、周瑜寡智，但是每次诸葛亮都会准确地猜中他的想法而对其进行伏击。曹操与诸葛亮的博弈，在于其互相在判断对方能够理解自己的想法的程度，而这个所谓的程度是个你用心理理论揣度我，而我用心理理论揣度你揣度我，你再用心理理论揣度你揣度我揣度你的无限过程。人工智能一定能无限运算下去，但这个无限过程在何处停止才真正标志人工智能的超凡智能。

三、人工智能体的道德规则（moral norm）

（一）人类的道德规则适用吗？

人工智能如果有道德能力，那么其基础便是道德规范或者规则（Malle, 2014）。而让其具有道德规范最为简单的方式便是将已有的人类道德规范加诸其身以对其进行出厂设置并规范其今后的学习与行为。那么人类规则是什么、如何适用便是首当其冲的问题。

人类的道德规则有哪些？通常情况下，人类倾向于将规则简约化，形成一整套所谓的价值观系统。价值观系统简单地在比较当两种情况相冲突时，人们看重（value）那一种，所谓价值亦即重量的含义。人类规范系统或曰价值观系统通常情况下有如下理论：第一，弗洛伦斯·克拉克宏（Florence Kluckhohn）和弗雷德·斯特罗德贝克（Fred Strodtbeek）提出了五种维度：人性（善、恶还是混合）、人与外在环境的关系（从属、主导还是与自然和谐相处）、人与他人的关系（等级、集体平权还是个性化的）、人的主要活动模式（存在、成为还是奋斗）、人的时间观念（看重过去、现在还是将来）（Kluckhohn & Strodtbeek, 1961）。第二，戈登·奥尔波特（Gordon W. Allport）将人类的价值观分为看重理论的、政治的、经济的、审美的、社会的还是宗教的（Vernon & Allport, 1931）。第三，米尔顿·罗克奇（Milton Rokeach）将价值观分为看重工具性价值观（有抱负、心胸宽广、有才能、快活、整洁、勇敢、助人、诚实、富于想象、独立、有理智、有逻辑性、钟情、顺从、有教养、负责任、自控、仁慈）还是终极性价值观（舒适生活、振奋生活、成就感、和平世界、美丽世界、平等、家庭保障、自由、幸福、内心平静、成熟的爱、国家安全、享乐、灵魂得到拯救、自尊、社会承认、真正友谊、智慧）（Rokeach, 1973）。第四，萨洛蒙·舒瓦茨（Shalom H. Schwartz）区分了十种价值观，并将其分为了自我超越（包含普世与慈善）、自我强化（包含权力与成就）、保守（包含传统、遵从与安全）与开放（包含自我导向、刺激与享乐）四类（Schwartz, 1992）。当然有些美德的分类通常情况下也可能被视为价值观的分类，如马丁·塞利格曼（Martin E. P. Seligman）和克里斯托弗·彼得森（Christopher Peterson）的六种美德模型：智慧与知识、勇气、仁慈、正义、节制、超越（Peterson & Seligman, 2004）；或者理查德·舒瓦德（Richard Shweder）的自治伦理、社群伦理和神性伦理（Shweder & Haidt, 1993）；亦或是乔纳森·海特（Jonathan Haidt）的伤害/关怀、公平/互惠、内群体/忠诚、权威/尊敬、纯净/圣洁五种道德基础（Graham, Nosek, Haidt, Iyer, Koleva, & Ditto, 2011）。实际上这些美德或者道德分类并非是看重何种价值的定义方式，将其看做价值观是存疑的（喻丰, 彭凯平, 董蕊, 柴方圆, 韩婷婷, 2013）。当然还有其他价值观的维度，比如文化价值观（个体主义—集体主义）、传统儒家价值观（如吃苦耐劳、服从权威、礼尚往来等）（彭凯平, 喻丰, 柏阳, 2011）。

这些价值观系统对于人工智能体来说适用吗？实际上这个问题非常简单，即不适用。因为人类价值观系统是一种抽象、尽量简约化的系统，而人工智能体所输入的必须是一种可以转化为变量或者代码表征的极其具体的行为规则。这种规则如果可以，最好表述为如果…那么…（if…then…）形式，这符合认知情感人格系统（CAPS）对于行为发生的定义（喻丰, 彭凯平, 韩婷婷, 柏阳, 柴方圆, 2012）。而如果将人类总结的抽象道德规范系统转化为具体的规则，这是一个演绎的过程，而具体需要演绎出多少特异化的规则，也就是说人工智能学习材料中需要包含究竟何种程度多样化的场景，这都需要探讨。通常情况下，人工智能科学家面对这样的问题会进行一个粗略的估计，如李飞飞等人在 ImageNet 中对于图片的人工标注

数量（320 万张或者 1500 万张）这只是一个粗略估计后看其识别率变化（Deng et al., 2009）。图片容易获得，而道德情境缺很难编制或者估计其数量。一个可能的途径是使用近期心理学对情境的分类模型（Rauthmann et al., 2014）重新对道德情境进行分类、演绎、编制、提取特征等过程来制作人工智能学习材料。

（二）如何解决道德规则冲突？

假使我们能够通过上述方式获得可能的具体的道德规范，那么这些道德规范是普遍适用的吗？实际上，道德规范当然有普遍的可能，但是即使是人类社会，也存在道德规范之间的相互冲突。对自动驾驶汽车道德规则的选取已然证明了文化甚至是性别差异（Awad et al., 2018）。有些文化中，女性被人看到会被认为是不道德的事情，被人强奸则父兄还会将其杀死，这对于我们看来不可思议（Fiske & Rai, 2014）。如果你是一个伴郎但是钱包被偷无法买票，你急着赶去婚礼送戒指，这个时候在火车站有一个偷他人钱包买票的机会，应不应该偷呢？美国人大多数觉得不应该偷，而印度人大多数觉得应该偷（Miller & Bersoff, 1992）。即使在同一种文化内，道德规则也可能产生冲突。如在狭路相逢被他人言语侮辱时，美国南方人比北方人的反应要表现出更多的攻击性，这是由于美国南方的荣誉文化所致（Cohen et al., 1996）。

但是我们试想，一个身处印度和身处美国的人工智能体，如果由同一家公司采用同一种基础的人工智能训练集和代码，那么对于类似上述是否该偷钱的情境，其表现必然是相同的。且不说其行为最终是偷亦或是不偷，这种相同合理吗？这个问题的回答应该对于不同的人工智能体来回答。笔者认为，如若人工智能体仅仅是完成某项单独任务，那么无须考虑类似文化差异问题。但如果人工智能体具有了社会规则，那么其本身便带有社会文化背景在进行行为，当然也包括道德行为。从这个意义上说，如果要让人类舒适，则在不同文化中使用的人工智能体必须使用所处文化的训练材料来进行学习，正如上述我们所说，哪怕是道德批判，中文和英文的词汇在意义和结构上都不尽相同。同时，处于不同文化中的人工智能体其先在道德假设也应符合其所处的文化。在上述例子中，美国机器人如果做出偷的行为，印度机器人如果做出不偷的决策，这似乎在其所处的文化中都不能称作道德。

此时我们便还能考虑另一种解决方案，即人工智能体或者机器人是否可以有双语甚至具有双语所代表的双文化规则？由于人工智能体现在还缺乏自我和意识，因此其双文化道德规则系统或者更多文化规则系统也许都是可以实现的，只是用两个训练集分别训练两个人工智能代码而存储在同一个实体中即可。但是如果人工智能在将来有了自我认同，那么这种情况是否还能出现？人工智能体是否能够双文化切换、是否能够同时容忍两种不同的道德价值存在这都是需要思考和探索，而现今并无答案的问题。

当然还有另一种可能性，即人工智能体无法同时容忍两种不同的道德价值存在，也就是说人工智能体拒绝道德相对主义存在的可能。那么在这样的情况下，人工智能体有可能促进一种所谓的普世价值，它促使文化融合，也促使全球化的状况加剧（Yu et al., 2016）。这是否是一件好事值得讨论，但也不失为一种可能性。

（三）人工智能体道德规则如何习得？

假使我们已然获得了大量可被表征的具体道德规范，也获得了解决不同文化社会情境下道德规范冲突的解决方案，那么我们是否就可能造出道德人工智能体了呢？事实上也并不那么简单。首先，我们必须有一个高度具体化和结构化的道德情境训练集，而这个训练集是能够避免不同社会文化情境中道德规范之间相互冲突的。以这个道德情境训练材料作为启示点的人工智能，还应被施加许多理论钦定的道德规则。这样的道德人工智能体在其诞生之初便有了很好的道德能力、尤其是指导其行为的规范。但这只是上述经由自上而下加工所产生的

最初道德人工智能体。

这些道德人工智能体不可避免地最终会接触形形色色的人与光怪陆离的社会现象，它可能会自下而上地习得新的规则。如何避免其习得人类并不喜欢的道德规则？似乎至少有三种途径可以选择：其一是，这些道德人工智能体已经在诞生之初便学习得特别顺畅，已然具有了明辨是非的能力（类似其知道选择何事可为、何事不可为，也知道为何），我们无需担心其之后的学习过程，因为它可以在学习过后弃恶存善。其二是，这些道德人工智能体在诞生之初已然学习了什么是好、什么是坏，但是人类依然对其存疑，我们限制其今后进行训练迭代的材料，制定一套专门挑选日后迭代训练材料的规则（类似让其选择何事可为、何事不可为，而不告诉其为何），让其主动（当然是规则所定）选择何合适的而向善的迭代学习材料，主动放弃可能会习得作恶的迭代学习材料。其三是，训练一种可能的自我控制能力（Baumeister, Vohs, & Tice, 2007），来约束人工智能体自己。从很多意义上来说，能够观察学习、自我调节和控制也许是人类最为伟大的、体现自己自由意志和能动性的行为了，正是这种特性造就了人类荣耀（喻丰, 彭凯平, 2018）。而第三种方式可能是人工智能真正成为人并超越人的存在方式，况且，这种自我调节还不能失败，这是必须要面对的问题。

四、人工智能体的道德理解（moral understanding）

（一）人工智能体会被人当成人吗？

人类有一种倾向，即对于无法理解的事物，我们倾向于使用能够理解的方式去扭曲这种事物，将其变为人类社会中存在的事物与现象来理解之。如果我们不知道人工智能是何物体，那么我们倾向于将其作为人看。这便是拟人化（Anthropomorphism）。它是一种将人类独有特质赋予非人实体的倾向性或形态（Epley, Waytz, & Cacioppo, 2007; 许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓, 2017）。人工智能通常情况下以拟人化的姿态出现，人们也倾向于将拟人化的机器看做人工智能体。譬如在街上随处可见所谓的拉面机器人，这种拉面机器实际上并非人工智能，而只是普通机器，但是由于其人形形象的存在，它通常情况下会被认为是人工智能。而真实的人工智能做饭或者炒菜的机器，由于其形态类似机器，而并无人形特征，在生活中常人却不会将其看做人工智能，知识匮乏者甚至无法理解这是人工智能的实体存在（喻丰, 许丽颖, In Press）。甚至普通人开始制造其所谓的人工智能时，均会选择拟人化的形态而开始（Broadbent, 2017）。

有人将人工智能看做一种新的物种，有人认为人工智能不足为惧，开玩笑言无非拔掉电源。这都是在指称人工智能的非生物有机体特性。但是这种有机体特性并不影响其拟人。人有将人工智能拟人化的倾向，人工智能也不可能不类似于人。它是由人类创造且在创造之初便模仿人类的智能体，即使其智能再过超群，它身上不可避免也有当初人类创造它并让其模仿人类的特征。至少在认识论层面上，这个时代我们将其看做一种介于机器与人之间的人性不够的人或者是具有某些人性的机器，因为它可能在人类知觉上具有了某些人性或者人类独特性特征（Haslam, 2006）。或许在今后的某个阶段，人工智能体会被看做超人，即在某些特性上它超过了人类。但无论是低人、非人还是超人，人工智能体都始终被人当做类似于人来看待，同样人类道德认知的方式，某种程度上在人工智能体上还起作用。

（二）人工智能体会改变人类关系吗？

拟人化的人工智能体，如机器人确实能够与人建立起社会联系（如陪伴机器人、性爱机器人与人之间必然能够建立社会心理联结），但是这种联系是否真的是社会关系是存疑的（Damiano & Dumouchel, 2018）。因为拟人化的机器人毕竟不是人，这种人与非人的关系只是拟人化外表赋予人与算法之间建立的关系，究其实质，这种关系是虚伪和欺骗甚至自我欺骗的关系。长久沉浸的与拟人化机器人建立起的关系，这甚至会反而影响人类真实的社会关

系（喻丰, 许丽颖, In Press）。

但是人工智能也许能够改变广义上人与物之间的关系。道德，尤其是中国传统道德最重要的就是体现人与人关系的调节功用。而人工智能体的出现，让人对待机器，尤其是对待具有道德立场的机器产生了极大的变化。普通民众实际上对普通机器的道德考量远逊于动物，因为前者缺乏人性和人类独特性，而后者只是缺乏人类独特性而已（Haslam, 2006）。至少，人们在普遍认识上认为在非人与人的直线上，机器不如动物。但是当人工智能兴起之后，这种关系是否会变化，关系之间的意义是否会随之变化，而其带来的道德后果会将如何，现在做结论还为时尚早。

（三）人工智能体能个性化吗？

一个有趣的事实是，人工智能在训练之后批量生产投入使用之前必然是普遍一般化的。举例来说，自动驾驶汽车在出厂时必然是一样的。但如果其具有学习功能，那么不同驾驶员如果在自动驾驶的同时还进行手动驾驶，那么不同驾驶员的驾驶习惯便会成为自动驾驶汽车自动学习的材料，这样自动驾驶汽车便可能在出厂后不就变得极具个性化，如果道路上均采用的是自动驾驶汽车的话，那么似乎除了人类无需付出驾驶劳动之外，道路状况与驾驶情境和之前并无本质区别。基于个体大数据的学习会使得人工智能变得高度个性化，这反而给人工智能体提供了一个信息茧房（喻丰, 彭凯平, 郑先隽, 2015）。

这种个性化是好事吗？理论上来说，现在的人工智能模拟的是人类的平均数，正如心理学研究的是人的心理过程和行为的平均数规律一样。以前述 ImageNet 为例，实际上人工智能进行学习的材料靠的是人工标注，而人工标注在这种图片识别的简单任务上都必然不大可能每个人一样，更遑论更加相对主义一些的道德了（喻丰, 韩婷婷, 2018）。值得注意的是，模拟的平均数实际上并不一定是普遍规律，因为平均数实际上并不能适用于每个个体，也许没有一个个体刚好是平均数，但个体们的集合造就了平均数。我们讨论的道德规范和道德能力并非是平均数规律，而是普遍规律。因此人工智能道德研究，存在一个从平均数规律向普遍规律过渡的阶段。平均数规律靠人工评定和计算便能够由科学家较为轻松地建立，而普遍规律则更需要人文社科学者的参与制定。所以从平均数规律向普遍规律的过渡正是人工智能从纯自然科学向包括人文社科在内的多学科交叉融合、以人为中心（human-centered）的过程。在这个过程中，可以预计的是以人类认知为中心、嵌入社会情境的社会心理学将发挥重要的作用。

参考文献

彭凯平 & 廖江群. (2009). 文化与归因的过程模式及概率模型探索. *中国社会科学*, (6), 31-40.

彭凯平, 喻丰, 柏阳. (2011). 实验伦理学：研究、挑战与贡献. *中国社会科学*, (6), 15–25.

王银春. (2018). 人工智能的道德判断及其伦理建议. *南京师大学报 (社会科学版)*, (4), 29-36.

邬家骅, 喻丰, & 许丽颖. (2016). 拟人化与道德责任. Unpublished manuscript.

许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓. (2017). 拟人化:从“它”到“他”. *心理科学进展, 25*(11), 1942–1954.

喻丰 & 韩婷婷. (2018). 有限道德客观主义的概率模型. *清华大学学报（哲学社会科学版）*, 33(3), 148-163.

喻丰 & 彭凯平. (2018). 文化从何而来. *科学通报, 63*(1), 32-37.

喻丰, 彭凯平, 董蕊, 柴方圆, & 韩婷婷. (2013). 道德人格研究：范式与分歧. *心理科学进展, 21*(12), 2235-2244.

喻丰, 彭凯平, 韩婷婷, 柏阳, & 柴方圆. (2012). 伦理美德的社会及人格心理学分析：道德特质的意义、困惑及解析. *清华大学学报（哲学社会科学版）*, *27*(4), 128-139.

喻丰, 彭凯平, 韩婷婷, 柴方圆, & 柏阳. (2011). 道德困境之困境：情与理的辩争. *心理科学进展, 19*(11),1702 - 1712.

喻丰, 彭凯平, 郑先隽. (2015). 大数据背景下的中国心理学：心理学的学科体系重构及特征.*科学通报, 60*(5-6), 520-533.

喻丰 & 许丽颖. (2018). 道德差序圈：中国人的道德结构. In Press at *南京师范大学学报（社会科学版）*.

喻丰 & 许丽颖. (2019). 道德责任归因：变与不变. In Press at *武汉科技大学学报（社会科学版）*.

喻丰 & 许丽颖. (In Press). 人工智能之拟人化. In Press at *西北师范大学学报（社会科学版）*.

Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, *18*(2), 103-115.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). *The Moral Machine experiment. Nature*, *563*(7729), 59-64.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*(6), 351-355.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*(1), 627–652.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. New York: Plenum press.

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology*, *70*(5), 945-959.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61-149.

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology, 9*, 468.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Li, F. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248-255). IEEE.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*, 864–886.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous Violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships*. Cambridge University Press.

Frimer, J. A., Tell, C. E., & Haidt, J. (2015). Liberals condemn sacrilege too: The harmless desecration of Cerro Torre. *Social Psychological and Personality Science*, *6*(8), 878-886.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366-385.

Gray, H. M., Gray, K. & Wegner, D. M. (2007). Dimensions of mind perception, *Science, 315,* 619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*, *23*(2), 101-124.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998-1002.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814-834.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog?. *Journal of Personality and Social Psychology*, *65*(4), 613-628.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social Psychology Review*, *10*(3), 252-264.

Kahneman, D., & Egan, P. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kluckhohn, F. R., & Strodtbeck, F. L. (1961). *Variations in Value Orientations.* Evanston, IL: Row, Peterson.

Malle, B. F. (2014). Moral competence in robots? In Seibt, J., Hakli, R., and Nørskov, M. (Eds.), *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014* (pp. 189-198). Amsterdam, Netherlands: IOS Press.

Malle, B. F., and Scheutz, M. (2014). Moral competence in social robots. In *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014* (pp. 30–35). Red Hook, NY: Curran Associates/IEEE Computer Society.

Miller, J. G., & Bersoff, D. M. (1992). Culture and moral judgment: How are conflicts between justice and interpersonal responsibilities resolved?. *Journal of Personality and Social Psychology*, *62*(4), 541-554.

Peterson, C., & Seligman, M. E. (2004). *Character Strengths and Virtues: A Handbook and Classification*. Oxford University Press.

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is human-like?: Decomposing robot human-like appearance using the Anthropomorphic roBOT (ABOT) Database. In *HRI '18: Proceedings of the Eleventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, Chicago, Illinois, USA. Piscataway, NJ: IEEE Press.

Rauthmann, J.F., Gallardo-Pujol, D.,Guillaume, E.M., Todd, E., Nave, C.S., Sherman, R.A., Ziegler, M., Jones, A.B., & Funder, D.C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107,* 677-718.

Rokeach, M. (1973). *The nature of human values*. New York: Free press.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10,* 173-220.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (pp. 165). San Diego, CA: Academic Press.

Shweder, R. A., & Haidt, J. (1993). The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological Science*, *4*(6), 360-365.

Singer, P. (2011). *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Vernon, P. E., & Allport, G. A. (1931). A test for personal values. The Journal of Abnormal and Social Psychology, 26, 231-248.

Voiklis, J., Cusimano, C., & Malle, B. F. (2014). A social-conceptual map of moral criticism. In P. Bello, M. Guarini, M. McShane, and B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1700-1705). AustinTX: Cognitive Science Society.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Yu, F., & Peng, K. (2016). Virtue, Continence, Incontinence and Vice: Making virtue judgments based on the judgment of thinking systems. *International Journal of Psychology, 51*(S1), 572.

Yu, F., Peng, T., Peng, K., Tang, S., Chen, C. S., Qian, X., Sun, P., Han, T., & Chai, F. (2016). Cultural value shifting in pronoun use. *Journal of Cross-Cultural Psychology, 47*(2), 310-316.